Test Time Impact of Redundancy Repair in Embedded Flash Memory

Paul Okino
Agilent Technologies

Abstract

Redundancy repair of high-density commodity Flash memory is an effective technique to improve per-wafer yield by trading-off increased die size and increased time at wafer-probe for the ATE system to analyze the failing bits and make the necessary repairs. In embedded Flash, where densities are typically much lower and test requirements more diverse, the benefit of redundancy repair is less certain. This paper discusses two key aspects of redundancy repair for embedded Flash memory blocks.

1. Introduction

Embedded memory of all kinds, and Flash in particular, are forecast to increase dramatically over the next two to three years, both in terms of the number of devices that incorporate Flash memory and the size of the embedded memory.

As these embedded blocks grow, the unique requirements of Flash memory test, coupled with the test requirements of the “host” device, may cause microcontroller and DSP suppliers to challenge their current assumptions about testing embedded memory, and provide a new opportunity to reduce cost-of-test.

2. Defect Density Considerations

Ignoring the increase in test time for redundancy repair, the optimum die area to allocate for redundant memory elements depends on the size of the array, defect density, and defect type.

Where densities for high volume commodity NOR-type Flash memories are typically in the range of 16 Mb to 256 Mb, embedded memory blocks in today’s “Flash-based” microcontroller and Digital Signal Processor (DSP) devices are typically three orders of magnitude smaller. Such small blocks of embedded Flash have an almost negligible effect on per-wafer yields, and redundant elements offer little or no benefit.

3. Test Time Effects

We began by ignoring the increase in test time for redundancy repair. Actually, there is a substantial increase in test time to analyze the location of the failing bits, decide the optimum allocation of redundant elements to repair those bits, and program the redundancy fuses, often referred to as “cams” by the Flash suppliers. So it would be a mistake to ignore the impact on test time when analyzing the effect of redundancy repair on cost-of-test.

This is especially true for Flash memory, whether commodity or embedded, where analysis of the failing bits and the repair strategy are implemented in a single-pass at wafer-probe test.

The fact that redundancy analysis and repair are performed on-line by the wafer-probe ATE system, as differentiated from the off-line repair methodology for DRAMs, is one of the two principal demands that have shaped modern Flash tester architecture—the other being the variability in the programming and erasing time for individual devices on the same wafer.

Tester architecture has a significant impact on the test time for Flash memory, especially when testing many die in parallel. An architecture that shares redundancy analysis hardware, or requires large data transfers between tester subsystems, generally suffers from much higher idle time.

4. Conclusions

As the density of embedded Flash memory increases, the need for redundancy repair will become more compelling. As the amount of redundancy repair increases, so too will test times – and a tester architecture focused on Flash, bundled with logic and analog test capabilities, will become an essential resource for testing a wide variety of Flash-based devices.